Cvičení ze statistiky - 8

Filip Děchtěrenko

Minule bylo..

- Dobrali jsme normální rozdělení
- Tyhle termíny by měly být známé:
  - Centrální limitní věta
  - Laplaceho věta (+ korekce na spojitost)
  - Konfidenční intervaly
  - Normální rozdělení
  - (Inferenční statistika)

Inferenční statistika

- Popsali jsme si deskriptivní statistiku a několik základních pravděpodobnostních modelů
- Inferenční statistika nám říká jak na základě vzorku (ten popíšeme pomocí deskriptivní statistiky) můžeme odvozovat parametry pravděpodobnostního modelu, ze kterého pocházejí data
- Můžeme na základě výběrových charakteristiky určit parametry modelu přesně?
- Ne -> Mluvíme o *odhadech* parametrů modelu

Odhady parametrů

- Můžeme odhadovat jednotlivé charakteristiky modelů, pak mluvíme o *bodových odhadech*
- Pro každý parametr máme ještě intervalový odhad (určuje, kde hledaný parametr nachází)
  - Vyberu jednoho člověka (věk 24 let). Výběrový průměr je 24 let. Co můžu říct o průměru celé populace?
    - Průměr celé populace je 24 let s pstí <0.001

Bodové odhady

- Jednotlivým výběrovým charakteristikám odpovídají parametry pravděpodobnostního modelu
- Nás budou zajímat tři
  - $\bar{x}$ je bodovým odhadem EX
  - s je bodovým odhadem σ
    s^2 je bodovým odhadem $varX$
  - p je bodovým odhadem π

Ukázka vztahu mezi vzorkem a populací

- Vygeneroval jsem 10 vzorků (každý má 10 pozorování) z rozdělení N(100,15) (ale to nevíme)

	var	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
1	1	10	104.1	16.43	109.0	106.25	13.34	68	123	55	-0.89	1.53	5.20
2	2	10	94.7	12.07	93.0	93.88	11.12	80	116	36	0.39	-0.44	3.82
3	3	10	92.9	18.24	94.0	93.88	25.95	64	114	50	-0.24	-1.37	5.77
4	4	10	101.5	15.77	102.5	101.25	17.79	79	126	47	0.01	-1.40	4.99
5	5	10	97.9	13.51	98.0	99.00	16.31	73	114	41	-0.32	-0.63	4.27
6	6	10	98.6	13.82	96.0	96.12	9.64	84	133	49	1.34	4.46	4.37
7	7	10	102.2	18.01	104.0	102.75	19.27	71	129	58	-0.24	-0.49	5.70
8	8	10	116.4	14.55	113.0	114.50	5.19	96	152	56	1.18	4.37	4.60
9	9	10	104.3	15.53	104.5	104.38	13.34	76	132	56	-0.03	0.56	4.91
10	10	10	107.9	17.40	104.5	106.00	9.64	88	143	55	0.86	0.79	5.50

- Celkový průměr je 102.05
- Co můžeme říct o tvrzení, že průměr je 100?

Intervalové odhady

- Chceme odhadnout, kde se vyskytuje odhadovaný parametr s danou pstí
- My budeme hledat intervaly spolehlivosti (konfidenční intervaly, CI) pouze pro π a μ
- Postup:
  - Stanovme požadovanou pst
    - Typicky se používá 99%,95% a 90%
  - Pro tyto hodnoty najdeme z hodnoty ze vzorečku P(-a<Z≤a)=2ϕ(a)-1
    - Dostaneme z hodnoty 2.576, 1.96 a 1.64
  - Upravíme (ukázka např. pro μ)

$$P(-z < Z < z) = P(-z < \frac{\overline{X} - \mu}{\sigma}\sqrt{n} < z) =$$
$$P(-z\sigma < (\overline{X} - \mu)\sqrt{n} < z\sigma) =$$
$$P(\overline{X} - z\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z\frac{\sigma}{\sqrt{n}})$$

  - A dosadíme
- Interval zapisujeme jako (D;H) kde D a H jsou dolní a horní mez

Další intervalové odhady

- Neznáme-li σ, použijeme t-rozdělení (později), případně do vzorečku dosadíme *s* namísto σ (nepřesné)
- Pro relativní četnost:

$$P(p - z\sqrt{\frac{\pi(1-\pi)}{n}} < \pi < p + z\sqrt{\frac{\pi(1-\pi)}{n}})$$

- Ve vzorečku nahradíme π pomocí *p*, tedy

$$P(p - z\sqrt{\frac{p(1-p)}{n}} < \pi < p + z\sqrt{\frac{p(1-p)}{n}})$$

Šíře intervalu

- Někdy nás zajímá, kolik potřebujeme lidí, abychom měli interval s daným rozsahem
  - Šíře klesá s rostoucí velikostí vzorku
  - Šíře roste s rostoucím koeficientem spolehlivosti
- Šíři CI určíme jako H-D, kde H a D jsou horní a dolní mez

Příklady

1. Opakovanými měřeními byla zjištěna tloušťka vlákna: 210, 217,209, 216, 216, 215, 220, 214, 213 (10^{-6} m). Je známo, že měření mají rozdělení N(μ, 25). Nalezněte 95% interval spolehlivosti pro μ.
   - Pro 95% CI je z hodnota 1.96, výběrový průměr je 214.4 a velikost vzorku je 9
   - CI tedy je (214.4-1.96·5/3; 214.4+1.96·5/3)=(211.1;217.7)
2. Jak by vypadal 95% CI, pokud bychom naměřili ještě 7 vláken 212,215,210,219,218,213 a 214?
3. Jak by vypadal 90% CI a 99% CI?
4. Kolik vláken bychom museli změřit, abychom měli 95% interval o šíři 2?
   - Dosadíme do vzorce H-D a dostaneme

     $$\left(\bar{X} + 1.96\frac{5}{\sqrt{n}}\right) - \left(\bar{X} - 1.96\frac{5}{\sqrt{n}}\right) < 2$$

     $1.96\frac{5}{\sqrt{n}}<1$

     n>96.04

     Potřebujeme tedy alespoň 97 vláken
5. Na dotazník ohledně preference zvířete odpovědělo 25 lidí. 7 z nich preferovalo kočku jako domácího mazlíčka. Určete 95% CI pro preferenci kočky v celém základním souboru
   - $p = \frac{7}{25}$

     95% CI tedy je (0.28-1.96$\sqrt{\frac{0.28(1-0.28)}{25}}$; 0.28-1.96$\sqrt{\frac{0.28(1-0.28)}{25}}$) $= (0.104; 0.456)$
   - Co je tady za metodologický problém?

Testování hypotéz

- Umíme popsat vzorek (deskriptivní statistika), známe základní pravděpodobnostní modely, nyní se naučíme, jakým způsobem z daného vzorku rozhodnout, zda mohou data pocházet z nějakého modelu.
- Metodu navrhl Fisher (p-hodnota) a upravili ji Neuman a Pearson
- Testováním hypotéz nemůžeme nic dokázat!! Pouze můžeme něco vyvrátit

Nulová a alternativní hypotéza

- Nulová hypotéza
  - H_0: tvrzení, které označuje stav, ve kterém se „nic neděje“
  - např. budu-li zkoumat, zda mince není falešná je H_0: mince je normální
- Alternativní hypotéza
  - H_1 (H_A): tvrzení, které vystihuje co se s daným problémem děje
  - např. H_A: Na minci padají častěji panny
- Hypotézy vyjadřujeme ve formě rovnic:

$$H_0: \pi = \pi_0$$
$$H_A: \pi > \pi_0$$

- Máme tři typy alternativních hypotéz: <,≠,>, používáme je podle výzkumné otázky (nebo zadání zkouškového příkladu☺)
- Pro <,> používáme jednostranný test, pro ≠ používáme dvoustranný test
- Pro testovanou hodnotu se podíváme, jak moc je odlehlá a podle toho zamítneme nebo přijmeme H_0

Hladina významnosti

- Musíme určit, s jakou pravděpodobností jsme udělali při našem soudu o hypotéze
- Můžeme udělat dvě chyby:
  - Chyba 1. druhu (α) : zamítneme H_0, přestože platí
  - Chyba 2. druhu (β) : přijmeme H_0, přestože neplatí
- Snažíme se minimalizovat pst obou chyb, ale univerzální postup neexistuje, protože spolu chyby souvisí
- Pravděpodobnost chyby prvního druhu (**hladinu významnosti**) stanovíme předem! Typicky se používají hodnoty 0.05, 0.01 a 0.001
- Pst 1- β se nazývá síla testu
- Chyby spolu souvisí, jediný způsob, jak snížit β (a nezvýšit α) je zvýšit velikost výběru

	Nezamítáme H_0	Zamítáme H_0
H_0 platí	1-α (OK)	α
H_0 neplatí	β	1- β (OK)

Příklad chyby 1. a druhého druhu

- Testování hypotéz se dá připodobnit soudu. Rozhodujeme o tom, zda odsoudit člověka ze zločinu. Možnosti jsou 4

	Vinen	Nevinen
Odsouzen	OK	Chyba 1. druhu
Zproštěn	Chyba 2. druhu	OK

- Snažíme se minimalizovat případy, že odsoudíme nevinného
- Nevinen, pokud se neprokáže jinak

Postup při testování hypotéz

1. Stanovíme si známé parametry
2. Stanovíme si nulovou a alternativní hypotézu
3. Stanovíme hladinu významnosti α a na jejím základě určíme kritickou hodnotu
4. Vypočítáme testovou statistiku (tohle je závislé na testu
5. Srovnáme testovou statistiku s kritickou hodnotou
6. Určíme p-hodnotu testu (=„Jestliže H_0 platí, jaká je pst, že získáme vypočítanou hodnotu nebo ještě neobvyklejší?“)

Druhy testů

- Celkem nás může potkat několik případů
- Budeme uvažovat, že data pocházejí z normálního rozdělení (jinak by se musely použít jiné metody)
- Zkoumáme μ základního souboru
  - Známe σ základního souboru
    - z-test
  - Neznáme σ základního souboru
    - t-test
- Porovnáváme μ_1 a μ_2 dvou základních souborů
  - Výběry na sobě závisí
    - Párový t-test
  - Výběry na sobě nezávisí
    - Dvouvýběrový t-test
- Zkoumáme-li σ základního souboru, používáme speciální testy pracující s χ^2rozdělením (nebudeme testovat) či F-testy (nebudeme testovat)

Z-test

- Používáme v případě, že známe σ základního souboru (to se nestává často)
- Výpočet testovací statistiky:

$$Z = \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n}$$

- Kde
  - $\bar{X}$ je výběrový průměr
  - μ_0 je teoretický průměr za případu, že by platila nulová hypotéza
  - σ je směrodatná odchylka základního souboru
  - n je velikost výběru

Rozhodnutí o přijetí nulové hypotézy

- Stanovíme si kritickou hodnotu *c* a spočítáme testovanou hodnotu *Z*
- Rozhodnutí o přijetí či zamítnutí závisí na tom, zda používáme jednostranný či oboustranný test
  - Jednostranný
    - Varianta „<„: zamítáme H_0, pokud je Z<c
    - Varianta „>„: zamítáme H_0, pokud je Z>c
  - Oboustranný
    - Zamítáme H_0, pokud je |Z|>c

Kritické hodnoty

- Pro dané hodnoty α jsou kritické hodnoty následovné

	Oboustranný	Jednostranný „<„	Jednostranný „>„
$\alpha = 0.01$	2.576	-2.33	2.33
$\alpha = 0.05$	1.96	-1.645	1.645

- U oboustranného testu budeme značit kritickou hodnotu N(0,1) rozdělení jako $u_{1-\frac{\alpha}{2}}$, tedy např. pro α=0.01->$u_{1-\frac{0.01}{2}} = u_{1-\frac{0.01}{2}} = u_{0.995}$=2.576
- U jednostranného testu budeme značit kritickou hodnotu N(0,1) rozdělení jako $u_{1-\alpha}$

Z-test pro alternativní znak

- Testovací statistika vypadá:

$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

- Kde
  - p je výběrová relativní četnost
  - π_0 je teoretická relativní četnost za předpokladu, že platí nulová hypotéza
  - n je velikost výběru

Příklad

1. Továrna na vlákna vyrábí vlákna o průměrné tloušťce 205 (10^{-6} m). Náhodně jsme vybrali několik vláken o tloušťkách 210, 217,209, 216, 216, 215, 220, 214, 213 (10^{-6} m). Je známo, že rozptyl výběru i základní populace je 25. Otestujte, zda náhodně vybraná vlákna jsou vyrobené v továrně
   - Parametry:
     $\bar{X}$=214.4
     μ_0=205
     n=9
     σ=5
   - Stanovíme H_0 a H_A
     $H_0: \mu = \mu_0$
     $H_A: \mu \neq \mu_0$ (oboustranný test)
   - Stanovíme hladinu významnosti a kritickou hodnotu
     α=0.05 -> $u_{1-\frac{0.05}{2}}$ =1.96 (oboustranný test)
   - Určíme Z
     Z=$\frac{214.4-200}{5}\sqrt{9}$=8.64
   - Oboustranný test -> Zamítáme H_0, pokud je |Z|>1.96
     To platí (8.64> 1.96), tedy zamítáme H_0 ve prospěch H_A
   - p-hodnota:
     P(|Z|>8.64)<0.001 (to je vidět), formálně:
     P(|Z|>8.64)=1-P(-8.64<Z<8.64]=1—(2ϕ(8.64)-1)=2-2=0<0.001

Příklad 2

- Na minci nám padlo 22 orlů ze 40 hodů. Preferuje tato mince orly? (řešte pro α=0.01)?
  - p=22/40=0.55, π_0=0.5, n=40
  - $H_0: \pi = \pi_0$
    $H_A: \pi > \pi_0$ (jednostranný test „>“)
  - α=0.01-> $u_{0.99}$=2.33
  - Z=$\frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$=$\frac{0.55-0.5}{\sqrt{\frac{0.5(1-0.5)}{40}}}$=0.63
  - Abychom zamítli, musí platit Z>2.33, což neplatí (0.63 < 2.33), Hypotézu H_0 tedy nezamítáme
  - P(Z>0.63)=1-P(Z≤0.63)=1-0.74=0.26 (a to je více než α=0.01)
- Kolik potřebujeme hodů, abychom tato nevyváženost (55%) byla významná?
  - Abychom mohli nulovou hypotézu zamítnout, musí platit, že Z>2.33, tedy
  - $\frac{0.55-0.5}{\sqrt{\frac{0.5(1-0.5)}{n}}}$>2.33
    0.05>2.33$\frac{0.5}{\sqrt{n}}$
    $\sqrt{n}$>23.3
    n>542.89
  - Potřebujeme tedy alespoň 543 hodů s 55% výběrovým poměrem, aby to bylo významné